
LightBits Lab storage appliance

“NVMe-over-TCP”

HyperScalers Pty Ltd.

Conducted at HyperScalers Proof of Concept (PoC) Lab 11th Mar 2020





Table of Contents

| | |
|---|---|
| 1. Executive Summary..... | 3 |
| 2. Introduction | 3 |
| 3. Test Environment..... | 4 |
| 4. Appliance architecture..... | 5 |
| 5. Benchmark tests | 7 |
| 6. LightField storage acceleration card | 8 |
| 7. Appliance accessibility | 9 |
| 8. Conclusion..... | 9 |



1. Executive Summary

At HyperScalers cloud enabled laboratory; we help customers design and perform proof of concept (PoC) on various cloud infrastructures; involving software and hardware platforms. The objective of this proof of concept is to benchmark the performance of LightBits Lab “NVMe-over-TCP” features on Hyper Scalers hardware (servers, switches) and evaluate their efficiency improvement after finetuning the appliance infrastructure building blocks.

The emergence of AI in real time applications and the need for bare metal infrastructure near the users, creates the need for the edge cloud. There is a general disregard to assess the storage demands at the edge, although it is an equally important tenant of edge computing. LightBits LightOS is the optimal storage disaggregation solution for the edge cloud. It is based on standard TCP/IP network and provides extremely low latency and high performance even with lower grade NVMe drives. LightOS for the edge supports inline compression and erasure coding and reduces the total cost of ownership by enabling edge clouds to leverage NVMe drives instead of HDDs that cannot meet many of the edge applications needs. Plus, it is a software-defined disaggregation solution that can run on any server.

Hyper Scalers works with multiple customers in telco, R&D labs and CSP; who use NVMe as an efficient storage layer. LightBits NVMe-over-TCP provides an efficient solution to scale accelerated storage across a scaled up environment.

2. Introduction

The NVMe data transfer has two performance parameters; one being flash drive and storage controller; other between the host and storage controller. In case of achieving data access as fast as NVMe, its less to do with improving flash drive speed, than improving the fabric or media over which data is transferred. Bringing NVMe’s massive parallelism to the data fabric promises to deliver huge performance improvements. To date FC and RDMA have been preferred media for NVMe drive access, the infrastructure to support this has kept some organizations out of the NVMe-over-Fibre market. To address this gap, the members of the NVMe.org consortium developed and published a new NVMe-oF standard (NVMe/TCP) using Ethernet LAN TCP datagrams as the transport medium.

There are numerous benefits of using NVMe-over-TCP:

- The standard uses TCP as the transport which is very common, well understood, and highly scalable protocol.
- Despite using Ethernet for connectivity, NVMe-over-TCP more closely resembles NVMe/FC because both use messages for their core communications, unlike RDMA-based protocols like RoCE that use memory semantics.
- There is a huge ecosystem of vendors in the TCP world, making major investments in improving its performance capabilities. Over the coming years, speeds are likely to increase significantly.
- Network design can have a huge impact on NVMe-over-TCP performance. In particular, the allocation of buffers needs to be “just right.” Too much buffering will add latency, and too little will result in drops and retransmission.
- NVMe over TCP is the newest fabric technology for NVMe; not much commercially available.

HyperScalers is an Australian registered company.

ABN - 83 600 687 223

ACN - 600 687 223

LightBits LightOS is a software-defined disaggregated storage solution uniquely tailored for edge storage. LightBits was among the first inventors of the NVMe-over-TCP standard, and LightOS SDS is the leading production-grade solution for NVMe-over-TCP. This solution means that now you can disaggregate your storage and get NVMe performance based on your choice of networks without any constraints.

The objective of this PoC is to qualify LightBits solution using Hyper Scalers hardware and capture performance metrics; to support underlying architecture.

3. Test Environment

The test environment consists of following hardware and software components:

| | |
|-----------------|--|
| Hardware | <p>Target Node – S5B QuantaGrid D52B-1U</p> <ul style="list-style-type: none"> ○ 2xIntel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz ○ 8x64GB RAM Samsung@ 2666 MHz DDR4 ○ 120 GB Intel SATA OS SSD ○ 4x 960GB HGST Ultrastar SN200 NVMe ○ 1xConnectX®-5 Dual-Port Adapter Supporting 100Gb/s Ethernet <p>Client Node – S2S QuantaPlex T41S-2U 4-Node (Only 2 Used)</p> <ul style="list-style-type: none"> ○ 2xIntel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz ○ 4x32GB RAM Micron@ 2133 MHz DDR4 ○ 120 GB Intel SATA OS SSD ○ 10G dual port SFP+ Mezzanine <p>Switches</p> <ul style="list-style-type: none"> ○ QuantaMesh BMS T4048-IX2 as leaf switch ○ QuantaMesh BMS T7032-IX1/IX1B as spine switch |
| Software | <ul style="list-style-type: none"> ○ CentOS Linux release 7.5.1804 ○ LightOS 1.3.5 |

Production enviroment built with [S5B T0](#)



S5B T0
Tier 0 Storage Server

Fastest Storage Ever
12 x 840K IOPVs per NVMe drive

- X 2** 40 CORES (CPU)
- X 16** 512GB (RAM)
- X 12** 184TB (NVMe SSD)
- X 2** 512GB SATADOM
- X 2** 4 X 100G (Network)

AUS home grown open manufacturer | Pre Configured

HYPERSCALERS

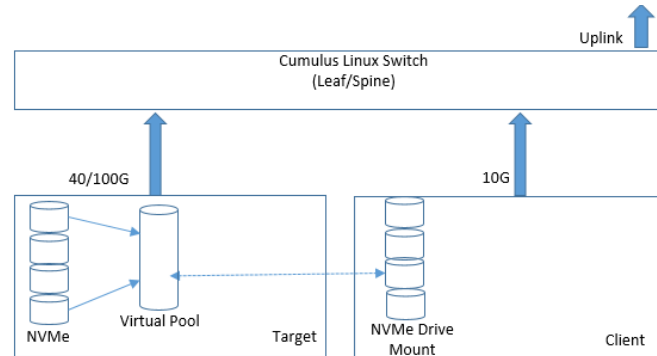
HyperScalers is an Australian registered company.

ABN - 83 600 687 223

ACN - 600 687 223

4. Appliance architecture

The diagram below depicts the system architecture for presenting NVMe-over-TCP.



The target node consists of 4 NVMe drives with 960GB each, connected over PCIe lane. The LightOS creates a virtualized pool of storage combining these drives and presents a single pool to the LightOS. The application can create multiple drives out of the pool and expose them over the high speed 100G data path to specific client. On the client node; the drive is presented as an NVMe drive; though via TCP network. The client OS can mount the drive and format it with specific file system. In the PoC, the mounted drive is executed with FIO and sysbench tools to evaluate their IO performance.

- List of NVMe drives connected to the target node

```
[root@localhost ~]# nvme list
Node          SN              Model          Namespace Usage
-----
/dev/nvme0n1  SDM00000D249   HUSMR7696BDP3Y1  1          960.20 GB / 960.20 GB
512 B + 0 B   KNGNP100
/dev/nvme1n1  SDM00000D27F   HUSMR7696BDP3Y1  1          960.20 GB / 960.20 GB
512 B + 0 B   KNGNP100
/dev/nvme2n1  SDM00000D259   HUSMR7696BDP3Y1  1          960.20 GB / 960.20 GB
512 B + 0 B   KNGNP100
/dev/nvme3n1  SDM00000D22F   HUSMR7696BDP3Y1  1          960.20 GB / 960.20 GB
512 B + 0 B   KNGNP100
```

- Creating a virtualized pool of storage and network resources on target node

```
[root@localhost ~]# lbctl get nodes
NAME      ID          LOGICAL USED      STATE      BOOT-STATE  EFFECTIVE CAPACITY  FREE      SUPPORTED CONFIGURATION  EC ENABLED  PHYSICAL USED
node1     586fd834-4266-4b65-a100-b1d6685d8d8c  Active      Enabled      2.4 TiB      1.4 TiB (1,589,040,319,693B)  1.0 TiB (1,099,511,627,776B)  [enp175s0f0 enp175s0f1 enp175s1f2]  true      false
```

The screen shot above shows an active node "node1" which represents the target node, with a virtualized pool of storage resource 2.4TB. The pool uses some space of disk for global FTL data; hence the pool size is slightly less than the physical space of each drive.

- Create volume and associate it with a client's UID

```
[root@localhost ~]# lbctl get volumes
NAME      ID          TYPE      CREATION-TIME  NODE-ID          CAPACITY  PHYSICAL USED  LOGICAL U
vol3      2dd11f38-d15e-4df3-b5f6-08ec72003274  NVMeOF    2020-03-05T06:41:48Z  586fd834-4266-4b65-a100-b1d6685d8d8c  3.0 TiB  0 B          0 B
vol1      b4850662-889c-4c89-a421-3ee6ff97eb97  NVMeOF    2020-03-03T12:38:19Z  586fd834-4266-4b65-a100-b1d6685d8d8c  1.0 TiB  1.0 TiB (1,099,511,627,776B)  1.0 TiB (1,099,511,627,776B)
vol2      e8a8f3db-8ac2-4532-9119-1545a0eb61ec  NVMeOF    2020-03-05T06:41:19Z  586fd834-4266-4b65-a100-b1d6685d8d8c  2.0 TiB  0 B          0 B
```

HyperScalers is an Australian registered company.

ABN - 83 600 687 223

ACN - 600 687 223

In this screenshot, there are 3 volumes created; named vol1/2/3; with varied sizes. The volumes are associated with ACL (Access control list); which points to the NQN (NVMe Qualified Name) of the client. This gives access to the client to attach the drive.

- On the client side, we needed to enable NVMe specific drivers for CENTOS; it would not be needed in UBUNTU releases, as they come with all modules to support NVMe over TCP.

```
[root@localhost ~]# lsmod | grep nvme
nvme_tcp          32768  0
nvme_fabrics     24576  1 nvme_tcp
nvme_core        102400  3 nvme_tcp,nvme_fabrics
```

- The drives created on target side are associated with the NQN of client node; hence to connect these drives, we needed to execute following command.

```
nvme connect-all -t tcp -s 4420 -a 10.0.10.1
```

The port used by NVMe is 4420 over tcp and the IP address points to the interface over which target exposes the drives for the clients to attach

- Once the drives are attached; its mounted with specific file system and can be used for executing performance benchmark tools like FIO and sysbench.

```
[root@localhost ~]# nvme list
Node          SN              Model          N
Namespace Usage
-----
/dev/nvme0n1  6b4fcbc214cd8809 LightBox       1
1.10 TB / 1.10 TB 4 KiB + 0 B 1.3
/dev/nvme0n2  6b4fcbc214cd8809 LightBox       2
2.20 TB / 2.20 TB 4 KiB + 0 B 1.3
```

In the attached screenshot, there are 2 drives connected through NVMe and are available as a regular drive on client node.

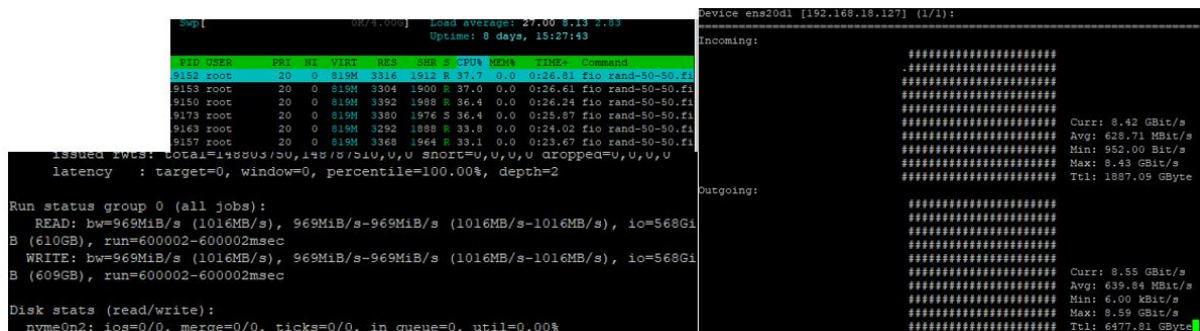
5. Benchmark tests

The PoC performed benchmark tests on mounted drive on client, using FIO and sysbench scripts. The reference benchmark are the tests conducted on Micron systems using similar architecture

<https://www.micron.com/about/blog/2020/march/nvme-over-tcp-proof-of-concept>

- FIO test

As part of FIO test the PoC executed random read & write tests and evaluated the CPU and network bandwidth utilizations in parallel.



The image contains two terminal screenshots. The left screenshot shows the output of the 'top' command, displaying system load averages and a list of processes. The 'fio' process is highlighted in green, showing it is using approximately 37% of the CPU. Below the 'top' output, the 'iostat' command is run, showing disk statistics for 'nvme0n2' with a utilization of 0.00%. The right screenshot shows the output of the 'nethogs' command for the network interface 'ens20d1', displaying incoming and outgoing traffic statistics. The incoming traffic shows a current rate of 8.42 Gbit/s, an average of 628.71 Mbit/s, and a total of 1887.09 GByte. The outgoing traffic shows a current rate of 8.55 Gbit/s, an average of 639.84 Mbit/s, and a total of 6477.81 GByte.

The screenshots above show the read and write performances on the attached drive to be around 1GB/s. The network bandwidth reaches around 10Gb/s on the client node, that is the peak. Also the CPU usage rises upto around 50%; hence with the BOM configuration; the script is able to stress the client node to maximum. The target node is connected with 40Gb/s CX-5 card, while the client node is on 10Gb/s data path, that is a bottleneck for the data transmission speed. Similarly a single client node is not sufficient to stress the data access to it's maximum in the PoC. As a next step of enhancement, Hyper Scalers plans to improve the BOM and number of clients attached to the target node.

- Sysbench test

Sysbench provides benchmarking capabilities to test CPU, memory, file I/O, mutex performance, and even MySQL on Linux. The PoC mounts attached NVMe drive to client filesystem and executes the sysbench tests.

```

Throughput:
  read, MiB/s:          6890.10
  written, MiB/s:       765.64

General statistics:
  total time:           60.0514s
  total number of events: 5262589

Latency (ms):
  min:                  0.00
  avg:                  0.63
  max:                  2233.11
  95th percentile:    3.82
  sum:                  3298869.04

Threads fairness:
  events (avg/stddev):  93974.8036/1378.51
  execution time (avg/stddev): 58.9084/0.03

real    1m0.076s
user    0m17.672s
sys     17m5.489s

```

The test performs random read/write with 128K blocksize and achieves read as 7GB/s and write as 765MB/s. The performance is considered better than FIO test, as sysbench does parallel 56 threads on mounted file system.

Commands used to create the test bed and execution are as mentioned below:

- `time sysbench --test=fileio --file-total-size=64G --file-test-mode=rndrw --num-threads=56 --file-block-size=128K prepare`
- `time sysbench --test=fileio --file-total-size=64G --file-test-mode=rndrw --max-time=60 --max-requests=0 --num-threads=56 --file-rw-ratio=9 --file-block-size=128K run`

The read/write ratio used is 9:1, that means 90% operations are performed as read and 10% as write. The number of parallel thread used is 56; the results prove that read performance on mounted file system is better with extensive parallelization of IOPs request. The PoC evaluated similar results using another in-memory software-defined-storage solution and finds the IOPs results better while using LightOS.

6. LightField storage acceleration card

Hyper Scalers would work with Lighbitslab to evaluate their data acceleration PCIe card LightField. The PCIe slot availability is already qualified on QCT box and it the PoC need to upgrade the BOM further. The LightField card is a PCIe storage acceleration option that seamlessly integrates with LightOS enabled systems. The card speeds LightOS's NVMe-over-TCP target and Global Flash Translation Layer (GFTL) with efficient hardware-based accelerated functions to improve overall system throughput, SSD utilization, and endurance that maximizes performance and extends LightOS TCO savings. LightField enabled systems experience improved utilization enabled by hardware-based at wire-speed data compaction and reduction with no performance degradation.





7. Appliance accessibility

The appliance can be made accessible to the customers using WAP DDNS “http://hyperscalers.asuscomm.com/”. Depending on the customer requirements; the administrator can open a port accessible via DDNS VPN.

8. Conclusion

Hyperscalers used its inhouse developed Digital-IP-Appliance Design Process along with a utility being the Appliance Optimiser Utility which together assists in productizing appliance(s) that enable providers with everything they need; to qualify LightBits Lab solution on its hardware.

The PoC executed NVMe-over-TCP software solution using LightbitOS environment and achieved the performance benchmark as per guidelines. The LightBits Lab comes with its own differentiations of creating remote low-latency pool of NVMe ssd.

| | | | |
|---|--|---|--|
|  <p>Bring Your Own Hardware for large IaaS/PaaS</p> <ul style="list-style-type: none"> • Software-defined solution • Standard NVMe SSDs support • Intel/AMD x86 architecture |  <p>Global FTL</p> <ul style="list-style-type: none"> • Thin Provisioning • Compression • RAID/EC • QoS services • Enhance Endurance and Latency • QLC support |  <p>No change in network (NVMe/TCP)</p> <ul style="list-style-type: none"> • Use vanilla TCP/IP network infrastructure - ubiquitous, simple & efficient • Run on standard Ethernet NICs |  <p>Don't touch clients (Target side solution)</p> <ul style="list-style-type: none"> • No proprietary client software • Standard NVMe/TCP client driver |
|---|--|---|--|

Hyper Scalers lab qualified these features to support customers who are already using NVMe supported storage pool. With this pre-qualified appliance, we target to reach out through common channel partners to potential users for NVMe-over-TCP. The performance benchmarks were compared with similar solutions qualified in the Hyper Scalers lab and totals IOPs with data read/wrote performance numbers “**Read:7GB/s, Write:1GB/s**” prove LightBits software defined storage solution to be high performing and optimized storage solution.